·图形图像处理·　　文章编号：1000-3428(2022)03-0244-09　　文献标志码：A　　中图分类号：TP391

基于学习主动中心轮廓模型的场景文本检测

谢斌红，秦耀龙，张英俊

（太原科技大学 计算机科学与技术学院，太原 030024）

摘　要：在场景文本检测领域，存在由于文本尺寸波动较大导致的小文本漏检、大文本欠检测和多尺度文本边界检测错误的情况。针对上述问题，提出一种基于学习主动中心轮廓模型的场景文本检测网络。在残差网络ResNet的基础上构建多尺度特征权重融合模型，对输入的场景文本图片进行多尺度特征提取和权重融合，并计算出最终的特征融合图，适应场景文本长宽比变化较大的情况。在此基础上，将融合后的特征图输入到学习主动中心轮廓模型预测文本框的中心点和边界，该模型为场景文本检测提供丰富先验知识，以解决多尺度文本检测框包含过多背景或部分包围文本造成的边界检测错误问题。在MSRA-TD500、IC13、IC15和IC17MLT数据集上的实验结果表明，该网络能够提高多尺度场景文本检测的准确率，其中在MSRA-TD50数据集上F-measure为0.83，相较于MSR方法提升1%，在IC13数据集上F-measure为0.91，相较于PixelLink网络提升2%，在IC15数据集上F-measure值为0.87，相较于PSENet网络提升1%，在IC17MLT数据集上F-measure值为0.74，相较于TridentNet网络提升1%。

关键词：场景文本检测；多尺度特征提取；权重融合；主动轮廓模型；学习主动中心轮廓模型

开放科学（资源服务）标志码（OSID）：

中文引用格式：谢斌红，秦耀龙，张英俊. 基于学习主动中心轮廓模型的场景文本检测[J]. 计算机工程，2022，48(3)：244-252，262.

英文引用格式：XIE B H，QIN Y L，ZHANG Y J. Scene text detection based on learning active center contour model[J]. Computer Engineering，2022，48(3)：244-252，262.

Scene Text Detection Based on Learning Active Center Contour Model

XIE Binhong，QIN Yaolong，ZHANG Yingjun

（School of Computer Science and Technology，Taiyuan University of Science and Technology，Taiyuan 030024，China）

【Abstract】In the field of scene text detection，there are several problems such as missing small text and insufficient precision for large text and multi-scale text boundary detection errors caused by large text size fluctuation. To solve the above problems，a scene text detection network based on a Learning Active Center Contour（LACC）model is proposed. First，the Multi-scale Feature Weight Fusion（MSWF）model is constructed on the basis of a Residual Network（ResNet）to extract multi-scale features and fuse weights of the input scene text images. Then the final feature fusion map is calculated to adapt to the situation where the aspect ratio of the scene text changes significantly. Finally the feature fusion map is then input into the LACC model to predict the center point and boundary of the text box，which provides rich prior knowledge for scene text detection to solve the problem of boundary detection errors caused by multi-scale text detection boxes containing too many backgrounds or partially enclosing text. Experimental results on MSRA-TD500，IC13，IC15 and IC17MLT datasets show that this network can improve the accuracy of text detection in multi-scale scenarios. The F-measure on MSRA-TD50 datasets is 0.83，which is 1% higher than the MSR method. The F-measure on IC13 datasets is 0.91，which is 2% higher than the PixelLink network. The F-measure on IC15 datasets is 0.87，which is 1% higher than PSENet. The F-measure on IC17MLT datasets is 0.74，which is 1% higher than TridentNet.

【Key words】scene text detection；multi-scale feature extraction；weight fusion；active contour model；Learning Active Center Contour（LACC）model

DOI：10. 19678/j. issn. 1000-3428. 0060828

基金项目：山西省重点研发计划（重点）高新领域项目（201703D111027）；山西省重点研发计划项目（201803D121048，201803D121055）。

作者简介：谢斌红（1971—），男，副教授、硕士，主研方向为图像处理、智能化软件工程、机器学习；秦耀龙，硕士研究生；张英俊，教授。

收稿日期：2021-02-07　　**修回日期：**2021-03-25　　**E-mail：**binhongxie@tyust.edu.cn

0　概述

场景文本检测有助于场景内容信息的获取、分析和理解，对于提高图像检索能力、工业自动化水平和场景理解能力等具有重要意义，可应用于自动驾驶、车牌票据识别、智能机器人、图片检索和大数据产业等场景。目前，场景文本检测已成为计算机视觉与模式识别、文档分析与识别领域的研究热点[1-2]。相较于通用目标检测，在同一张或不同自然场景图片中文本尺度变化较大，最大文本与最小文本之间可以相差近230倍[3]。因此，多尺度场景文本检测网络应运而生，多尺度场景文本检测网络通过多尺度多形式的特征提取和融合，以适应场景文本尺度的多变性，对于场景文本检测的工业化应用具有十分重要的意义。

早期多尺度场景文本检测网络采用传统机器学习的方法，通过传统的图像处理方法和人工设计的特征检测场景文本。如文献[4]通过提取最大稳定极值区域（MSER）找出候选字母，根据Hu矩特征刻画候选字母的几何特征；然后通过单链接聚类得到候选文本，最后引入共生纹理特征筛选文本区域，该算法对文本尺度变化有一定的适应性。文献[5]采用基于方向预分类的Gabor小波变换特征提取方法，利用Gabor函数良好的频率选择性和方向选择性，同时考虑到笔画相对位置的偏移，对笔画变形和低分辨率字符具有较好的适应性。文献[6]引入笔画宽度变化（SWT）算法处理场景图片提取不同尺度和不同方向的文本候选区，采用手工特征和随机森林（Random Forest，RF）算法过滤非文本区域，利用文本间的相似性连接成文本行。文献[7]采用多尺度滑动窗口模型，针对文本的局部特征提出一种基于文本部件的树形结构，该算法能很好地适应文本尺度的多变性。上述方法虽然在多尺度场景文本检测领域取得了不错的效果，但是传统的机器学习方法在特征提取方面仍有许多不足：由于场景文本检测的复杂性，人工设计特征难度高，需要消耗大量的时间和人力，成本较高；人工设计的特征会引入人为因素，可能造成文本特征的缺失甚至引入错误特征，检测精度不高。

近年来，随着深度学习网络的迅猛发展，涌现出一系列多尺度场景文本检测网络，其中典型方法是基于金字塔网络，根据其网络结构可分为单向金字塔网络和双向金字塔网络。

基于单向金字塔网络的方法[8-11]通常只有自下而上的特征提取过程，即对原始输入图像通过卷积、池化等操作进行特征提取，在不同层的特征图或者融合后的特征图上进行文本检测。根据网络输入不同大致可分为两类：一类为单向特征金字塔网络，其输入为单一尺度场景图片，在同一图片不同层的特征图或者不同层融合后的特征图上进行文本检测；另一类为单向图片金字塔网络，该类方法输入不同尺度图片，在不同图片的不同尺度特征图上进行检测。如CTPN[8]为解决文本长度变化非常剧烈的问题，采用单向特征金字塔网络，通过设定相同宽度不同高度的垂直文本候选框，使用长短期双向记忆模型处理文字建议序列，进而计算多尺度文字区域外接框与置信度。R2CNN[9]使用Faster R-CNN网络提取特征，利用不同尺寸卷积核的ROI Pooling处理特征图，从而计算出文本目标的矩形包围框检测多尺度场景文本。TextBoxss[10]基于SSD框架，根据不同卷积层的多尺度特征检测不同尺寸文本，通过设定不同纵横比的默认文本候选框，提高了不同尺寸文本的检测准确率。SSTD[11]基于SSD框架，借鉴了GoogleNet[12]中的Inception模块，使用HIM（Hierarchical Inception Module）融合卷积特征以提高模型的性能。

基于单向金字塔结构网络虽然在多尺度场景文本检测领域取得了较好的性能，但是单向特征金字塔网络只包含自下而上的特征提取过程，可以提取到丰富高层语义特征，却忽略了低层特征图包含的文本边界特征信息，造成文本边界检测不准确；而单向图片金字塔网络虽然通过对不同尺度图片进行特征提取，有利于检测不同尺度文本，但是增加了额外的计算开销。因此，针对单向金字塔网络的不足，研究人员提出了基于双向金字塔结构的多尺度场景文本检测网络。

基于双向金字塔网络的方法包括自下而上和自上而下两个特征提取过程，首先在自上而下过程中融合同层次自下而上的特征图，最后在不同尺度特征图或融合后的特征图上进行检测，该类方法可以有效利用低层特征分辨率高和高层特征语义信息丰富的特点，其中低层特征语义信息少但分辨率高，有利于文本目标的定位，而高层特征分辨率低但包含语义信息丰富，有利于文本目标和非文本目标的分类。典型方法为基于FPN[13]（Feature Pyramid Network）结构，如PSENet[14]引入FPN结构，首先将文本区域收缩划分为多级中心区域，然后进行像素级分类和预测多级中心区域，最后使用广度优先搜索算法逐级扩展为不同的文本区域，可以有效地检测相距较近和尺度多变的文本实例。MSR[15]网络使用FPN结构，输入多个尺度原始图像进行特征提取并在相同层对不同尺度图片的特征进行融合，在最后融合后的特征图上检测多尺度的文本。CCTN[16]网络基于VGG框架，首先将文本划分为文本区域和文本中心线区域，然后通过先粗略分类后精细分割的方式检测多尺度文本。EAST[17]则基于PVANet，通过上采样融合不同网络层的特征计算出融合特征图，在融合特征图的基础上回归文本包围框的相关属性来检测场景文本。PixelLink[18]采用基于VGG-16的双向特征金字塔网络，通过预测像素类别和像素在空间特征上的连通性以区分不同文本。

基于双向金字塔网络的方法虽然同时融合高层语义特征和低层包含的边界特征，但其在自下而上的深层特征提取过程中，会通过下采样来减小特征图的分辨率以增大感受野来提取高层的语义特征，因此存在以下不足：大文本边界回归弱，虽然较深的特征图有利于预测大文本，但随着特征图分辨率的减小和网络层数的加深，大文本的边界信息也会逐渐减少，因而不利于大文本的边界回归；小文本语义信息丢失。随着特征图分辨率的减小，小文本的语义特征会逐渐减少甚至丢失，导致出现漏检情况；文本边界检测错误，随着大文本边界信息的减少和小文本语义特征的减弱，造成文本检测框包含过多背景或部分包围文本造成的边界检测识别错误，虽然该类方法在自上而下的过程中融合了同层次的自下而上的浅层特征图以增强低层边界特征，但是对于大文本而言，浅层特征图的语义信息弱，没有能力检测大文本，而对于小文本而言，由于其语义特征已经丢失，即使融合特征，检测效果也不会有明显提升。

针对金字塔结构由于下采样减小特征图分辨率而造成的性能次优情况，本文采用多尺度特征权重融合（Multi-Scale Feature Weight Fusion，MSWF）模型在3个分支上进行多尺度特征提取，以兼顾高层语义特征和低层边界特征。由于不同分支特征之间的不一致性，本文在3个分支加入可学习的权重，并针对多尺度场景文本边界检测错误的问题，提出学习主动中心轮廓（Learning Active Center Contour，LACC）模型用于多尺度场景文本边界检测。

1 基于LACC模型的场景文本检测网络

1.1 场景文本检测算法框架

本文提出的基于学习主动中心轮廓模型的场景文本检测网络结构如图1所示。本文网络使用ResNet[19]作为主干网，首先在其基础上构建多尺度特征权重融合（MSWF）模型，在3个不同分支上进行多尺度的特征提取和权重融合，然后借鉴FPN结构提取自上而下特征图，接着使用融合函数C对自上而下各层特征图进一步融合计算出最终特征图F，将融合后的特征图输入学习主动中心轮廓模型中进行中心点的定位和边界框的回归，最后得出预测结果。

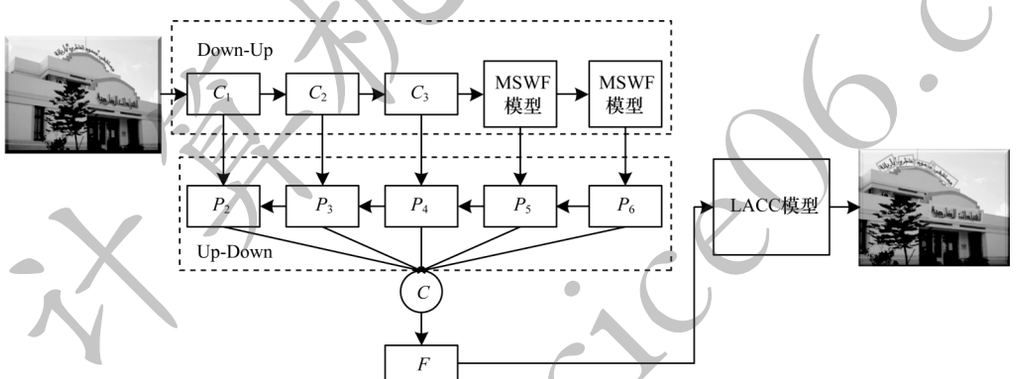


图1 基于学习主动中心轮廓模型的场景文本检测网络

Fig.1 Scene text detection network based on learning active center contour model

1.2 文本检测网络特征提取

本文网络特征提取过程如图1所示，首先将增强后的场景文本图片输入到自下而上的主干网中，依次计算出3个不同尺度的特征图(C_1,C_2,C_3)，然后将C_3输入到2个多尺度特征权重融合模型中，自上而下特征提取过程与FPN[13]相同，最后获得5个256通道的特征图(P_2,P_3,P_4,P_5,P_6)，并使用连接函数C进一步融合不同层次的特征图，得到具有1 280通道的特征图F，该函数定义如下：

$$F=C(P_2,P_3,P_4,P_5,P_6)=P_2\|\mathrm{Up}(P_3)\|\mathrm{Up}(P_4)\|\mathrm{Up}(P_5)\|\mathrm{Up}(P_6) \tag{1}$$

其中："‖"为连接（concatenation）操作；Up()为上采样操作。特征图F被输入到可变形卷积网络[20]（DCNv2）中进一步提取特征并且将通道数降为64通道，然后将其输入到学习主动中心轮廓模型中进行中心点的定位和文本边界的回归，并计算出检测结果。

1.3 多尺度特征权重融合模型

多尺度特征权重融合模型结构如图2所示，将图1所示特征图C_3输入到多尺度特征权重融合模型中，该模型包含3个分支（Branch1，Branch2，Branch3），各分支处理流程相同。以Branch1分支为例，特征图C_3经过空洞卷积块计算空洞卷积特征图D_1，该分支分为上下两层，上层通过权重计算模块计算出该分支的权重特征图W_1，下层输出卷积特征图D_1，此时Branch1分支输出特征图B_1，满足：

$$B_1=W_1\odot D_1 \tag{2}$$

其中：$\odot$为Hadamard乘积，按照上述流程依次计算出各分支输出特征图(B_1,B_2,B_3)，则MSWF Mode输出特征图MF，满足：

$$M_{\mathrm{MF}}=\mathrm{cat}(B_1,B_2,B_3) \tag{3}$$

相较于三叉戟网络[21]（Trident Network，TridentNet），本文网络不同之处在于着重对3个分支的融合过程

进行改进，TridentNet三分支卷积核参数权值共享，首先对训练目标进行尺度划分，然后根据划分结果选择分支进行训练，最后使用第二分支进行推理，并经过NMS后处理输出推理结果。

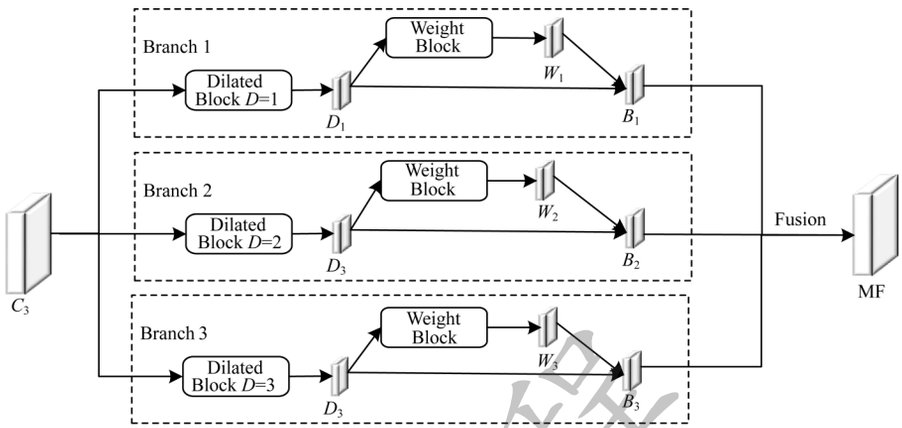


图2　多尺度特征权重融合模型示意图

Fig.2　Schematic diagram of multi-scale feature weight fusion model

因此，MSWF Model与TridentNet存在以下不同之处：1）TridentNet对训练目标进行尺度划分来选择训练分支，三分支卷积核参数权值共享，而本文采用三分支并行训练，分别计算三分支权重；2）TridentNet通过第二分支进行检测，本文在权重融合后的特征图上检测文本；3）TridentNet选择某个分支进行训练和检测属于硬注意力的一种，本文对三分支的特征基于权重融合属于软注意力。

MSWF模型各分支的空洞卷积[22]分别使用了不同空洞率（Dilate=1，Dilate=2，Dilate=3），采用三分支结构是为了在多个分支提取多种尺度特征，以适应场景文本的多尺度变化。相同分支则在同一尺度采用空洞卷积，既可以保持分辨率不变，又可以扩大感受野，以此取代特征金字塔下采样提取文本特征的方法，提高了网络对于大、小文本的检测性能。

图2中权重计算模块结构如图3所示，D_1、D_2、D_3分别经过1×1Conv层、BN（Batch Normalization）层和ReLU（Rectified Linear Units）层生成特征图（W_1_t, W_2_t, W_3_t），然后使用cat（concatnate）函数将其拼接为特征图W_t，再经过SoftMax函数作归一化处理得到可学习的权重特征图（W_1, W_2, W_3）。

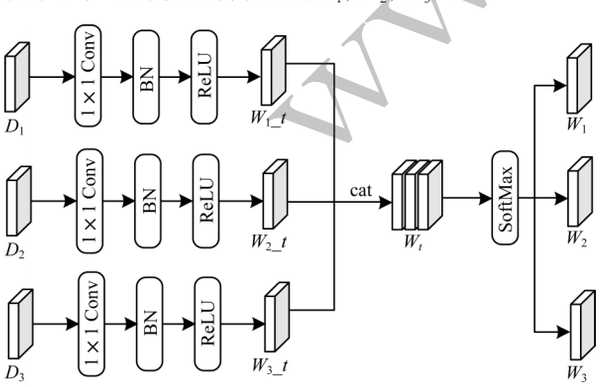


图3　权重计算模块结构

Fig.3　Structure of the weight calculation block

在具体计算权重时，令$\boldsymbol{x}_{ij}^{l}$为$l(l\in[1,3])$分支经过空洞卷积模块后产生的特征图（D_1, D_2, D_3）在(i,j)位置的特征向量，则有：

$$\boldsymbol{y}_{ij}=\boldsymbol{\alpha}_{ij}\cdot\boldsymbol{x}_{ij}^{1}+\boldsymbol{\beta}_{ij}\cdot\boldsymbol{x}_{ij}^{2}+\boldsymbol{\gamma}_{ij}\cdot\boldsymbol{x}_{ij}^{3} \tag{4}$$

其中：$\boldsymbol{y}_{ij}$表示多尺度特征权重融合模型输出的特征图MF在(i,j)位置的特征向量；$\boldsymbol{\alpha}_{ij}$、$\boldsymbol{\beta}_{ij}$、$\boldsymbol{\gamma}_{ij}$分别表示三分支权重特征图（W_1, W_2, W_3）在(i,j)位置的特征向量。受文献[23]的启发，本文令$\boldsymbol{\alpha}_{ij}+\boldsymbol{\beta}_{ij}+\boldsymbol{\gamma}_{ij}=1$，且$\boldsymbol{\alpha}_{ij},\boldsymbol{\beta}_{ij},\boldsymbol{\gamma}_{ij}\in[0,1]$，如式（5）所示：

$$\boldsymbol{\alpha}_{ij}=\frac{e^{\lambda_{\alpha_{ij}}}}{e^{\lambda_{\alpha_{ij}}}+e^{\lambda_{\beta_{ij}}}+e^{\lambda_{\gamma_{ij}}}} \tag{5}$$

其中：$\boldsymbol{\alpha}_{ij}$、$\boldsymbol{\beta}_{ij}$、$\boldsymbol{\gamma}_{ij}$分别由以$\lambda_{\alpha_{ij}}$、$\lambda_{\beta_{ij}}$、$\lambda_{\gamma_{ij}}$作为控制参数的softmax函数计算得出。

1.4　MSWF算法

算法1　MSWF算法

输入　ResNet50第三层输出卷积特征图C_3

输出　多尺度特征权重融合特征图MF

步骤1　将C_3输入模型的第$i(i\in[1,3])$分支中。

步骤2　使用Dilated Block$_i$空洞卷积块对C_3进行特征提取，得到特征图D_i。

步骤3　将D_i依次输入图3中Conv、BN、ReLu层计算出权重特征图W_i_t。

步骤4　重复步骤2和步骤3，依次计算出卷积特征图（D_1, D_2, D_3）。

步骤5　重复步骤4依次计算出权重特征图（W_1_t, W_2_t, W_3_t）。

步骤6　使用cat函数拼接（W_1_t, W_2_t, W_3_t），计算出融合权重特征图W_t。

步骤7　将W_t输入到softmax函数归一化处理，计算出各分支权重特征图（W_1, W_2, W_3）。

步骤8　计算权重特征图（W_1, W_2, W_3）与卷积特征图（D_1, D_2, D_3）的Hadamard乘积，结果为各分支输

出特征图(B_1,B_2,B_3)。

步骤9 使用cat函数拼接特征图(B_1,B_2,B_3),计算出MSWF Model输出特征图MF。

2 损失函数

2.1 主动轮廓模型

在众多分割方法中,基于主动轮廓模型(ACM)或其变形模型是图像分割中应用最广泛的方法之一,取得了较好的性能。本文提出的学习主动中心轮廓模型即是受主动轮廓模型的启发。1988年,KASS等[24]提出主动轮廓模型,将图像分割问题转换为求解能量泛函最小值问题,为图像分割提供一种全新的思路。该模型的主要原理是通过构造能量泛函,在能量函数最小值驱动下,使用基于偏微分的方法最小化能量函数,使轮廓线朝着目标边界的方向不断演进,最终分割出目标。但由于实际图像的背景是不均匀的,并且背景和目标的对比度往往比较低,仅依靠能量函数的最小值无法准确分割出目标,因此Chan-Vese模型将曲线所围的面积和曲线长度作为能量项引入到能量函数中。在过去若干年里,研究人员提出了很多基于ACM的模型,如无边缘主动轮廓模型(ACWE)和BRESSON等[25]提出的快速全局最小化主动轮廓模型(FGM-ACM)。

ACWE模型的能量最小化问题可以表述为:

$$\min_{\Omega_c,c_1,c_2}\left\{E_1^{\mathrm{ACWE}}(\Omega_c,c_1,c_2,\lambda)=\int_0^{\mathrm{Length}(C)}ds+\lambda\int_{\Omega}(c_1-f(x))^2\,\mathrm{d}x+\lambda\int_{\Omega/\Omega_c}(c_2-f(x))^2\,\mathrm{d}x\right\} \tag{6}$$

其中:ds是欧几里得长度;C是曲线的长度;$f(x)$是待分割图像;Ω_c是图像$f(x)$在图像域Ω上的闭合子集;c_1是内部区域的平均灰度;c_2是外部区域的平均灰度;λ是用于控制正则化过程中c_1、c_2之间平衡的参数($\lambda\succ 0$)。

2.2 学习主动中心轮廓模型

虽然基于主动轮廓模型的图像分割取得了很好的效果,但还存在以下不足:1)采用无监督的方法不需要从训练数据中学习属性,因此它们很难处理噪声和遮挡;2)有许多参数是依据经验设定的;3)多数方法都不能对自然场景图片进行鲁棒分割。显然,基于主动轮廓模型的方法不适用于有监督的机器学习来处理标记图像,而且大多数基于深度学习的自然场景文本检测方法缺乏整合目标先验知识的机制。因此,有必要将基于深度学习的场景文本检测与基于主动轮廓模型的方法结合起来,以便后者能够提供足够的先验知识,以提高模型对于场景文本边界的检测性能。

本文受主动轮廓模型ACWE模型能量最小化问题的启发,提出一个整合了中心点、文本区域和文本检测框长度信息的可用于深度神经网络学习的新损失函数,将LACC Mode的先验知识用于神经网络训练,解决多尺度场景文本检测框轮廓线能量全局最小化问题,进而精确检测文本边界。具体的损失函数如式(7)所示:

$$\mathrm{Loss}_{\mathrm{LACC}}=\lambda_c\mathrm{center}+\lambda_l\mathrm{length}+\lambda_r\mathrm{region} \tag{7}$$

其中:

$$c_{\mathrm{center}}=\frac{-1}{N}\sum_{xyc}\begin{cases}(1-\hat{Y}_{xyc})^{\alpha}\log_a(\hat{Y}_{xyc}), Y_{xyc}=1\\(1-Y_{xyc})^{\beta}\log_a(\hat{Y}_{xyc})^{\alpha},\text{其他}\\\log_a(1-\hat{Y}_{xyc})\end{cases}$$

$$l_{\mathrm{length}}=\sum_{\Omega}^{i=1,j=1}\sqrt{\left|(\nabla u_{x_{i,j}})^2+(\nabla u_{y_{i,j}})^2\right|+\varepsilon}$$

$$r_{\mathrm{region}}=\left|\sum_{\Omega}^{i=1,j=1}u_{i,j}(c_1-v_{i,j})^2\right|+\left|\sum_{\Omega}^{i=1,j=1}(1-u_{i,j})(c_2-v_{i,j})^2\right| \tag{8}$$

其中:center中心点表示矩形文本框的中心点,采用Focal Loss[26]损失函数;$\hat{Y}$为文献[27]中的关键点热力图;α和β是文献[27]中的超参数,根据文献[27]设置$\alpha=2,\beta=4$;N是输入图像中的关键点个数;length表示矩形文本框的周长;region表示矩形文本框的面积;$v,\mu\in[0,1]^{m\times n}$分别表示文本框标注值和预测值;$u_{x_{i,j}}$和$u_{y_{i,j}}$中的$x_{i,j}$和$y_{i,j}$分别表示本文网络在特征图$(i,j)$位置的水平方向和垂直方向;$\varepsilon(\varepsilon>0)$是为了防止平方根为零而添加的参数,在训练时,设$\varepsilon$为一个极小的正数即可;$c_1$和$c_2$分别表示内部和外部能量。

c_1和c_2的定义如下:

$$c_1=\int v\cdot\mu\mathrm{d}x\Big/\int\mu\mathrm{d}x$$

$$c_2=\int v\cdot(1-\mu)\mathrm{d}x\Big/\int(1-\mu)\mathrm{d}x \tag{9}$$

综上,本文提出一种新损失函数,考虑了文本的中心点和边界轮廓的长度以及文本区域的拟合度,具有以下优点:1)将无监督AC Model能量最小化问题转化为有监督的深度学习损失函数最小化问题;2)将本文网络提取特征和AC Model的先验知识相结合,解决了AC Model过于依赖人工特征设计、鲁棒性差和深度学习缺乏足够先验知识的问题;3)将AC Model基于图像像素信息检测方式转化为基于深度学习卷积特征图像素信息检测方式。

2.3 学习主动中心轮廓模型算法

算法2 LACC算法

输入 本文网络输出特征图F,文本中心点标注值Y_{xyc},文本框标注值v,超参数$\varepsilon(\varepsilon>0)$

输出 网络损失$\mathrm{Loss}_{\mathrm{LACC}}$

步骤1 对特征图F进行解码(decode),得出文本中心点预测值$\hat{Y}_{xyc}$和文本框预测值μ。

步骤2　初始化 $\mathrm{Loss}_{\mathrm{ACC}}=0$，$\varepsilon=10^{-8}$，center=0，length=0，region=0，$c_1=1$，$c_2=0$，$\lambda_c=1$，$\lambda_l=0.5$，$\lambda_r=0.5$。

步骤3　判断 Y_{xyc} 是否满足 $Y_{xyc}=1$，如果满足则转步骤4，否则转步骤5。

步骤4　如果 $Y_{xyc}=1$，则中心损失计算如下：

$$c_{\mathrm{center}}=\frac{-1}{N}\sum_{xyc}(1-\hat{Y}_{xyc})^{\alpha}\log_a(\hat{Y}_{xyc})$$

其中：$\alpha=2$；$\beta=4$；N 是输入图像中的关键点个数。

步骤5　如果 $Y_{xyc}\neq 1$，则中心损失计算如下：

$$c_{\mathrm{center}}=\frac{-1}{N}\sum_{xyc}(1-Y_{xyc})^{\beta}\log_a(\hat{Y}_{xyc})^{\alpha}\log_a(1-\hat{Y}_{xyc})$$

步骤6　计算文本框的长度损失length，具体公式见式(8)。

步骤7　计算文本框区域损失region，具体公式见式(8)。

c_1 和 c_2 分别表示内部和外部能量，定义见式(9)。

步骤8　计算网络损失函数 $\mathrm{Loss}_{\mathrm{LACC}}$，具体公式见式(7)。

步骤9　得到学习主动中心轮廓模型算法损失 $\mathrm{Loss}_{\mathrm{LACC}}$。

3　实验结果与分析

3.1　数据集

MSRA-TD500[6]是一个文本尺度变化较大的中英文数据集，共包含500张图片，其中300张用于训练，200张用于测试。

ICDAR-2013(IC13)[28]是一种常用的多方向英文场景文本检测数据集，共包含509张图片，其中258张图片用于训练，251张图片用于测试。文本区域是由四边形的左上、右下2个顶点标注。

ICDAR-2015(IC15)[29]是一种常用的多方向英文文本检测数据集，共包含1 500张图片，其中1 000张图片用于训练，500张图片用于测试。文本区域是由四边形的4个顶点标注。

ICDAR-2017MLT(IC17-MLT)[30]是一个大规模的多方向多语言场景文本数据集，包括7 200张训练图片、1 800张验证图片和9 000张测试图片，由9种自然语言组成，文本区域由四边形的4个顶点标注。

3.2　实验参数

本文使用ResNet50[19]在ImageNet[31]上的预训练模型，并在其基础上构建多尺度特征权重融合模型(MSWF Model)作为网络的主干网，所有网络都采用Adam[32]优化器进行训练。首先在IC17-MLT数据集上进行训练，得出IC17-MLT上的训练模型并进行测试，然后加载该训练模型为预训练模型，在MSRA-TD500、IC13和IC15数据上分别继续训练网络并进行测试。本文实验在GTX1080Ti×2个GPU上进行批量大小为8的270K次迭代。初始学习率设置为 1×10^{-4}，在90K次和180K次迭代时将学习率分别调整为 1×10^{-5} 和 1×10^{-6}。

在训练时，忽略数据集中标注为“不用关注”的模糊文本区域。本文根据实验结果，将损失函数的权重系数设定为：$\lambda_c=1$，$\lambda_l=0.5$，$\lambda_r=0.5$。采用以下方法对训练集的数据进行增强：1)图像按比例在[0.6, 1.4]之间以步长为0.1进行随机缩放；2)图像在[−10°, 10°]范围内随机进行水平翻转和旋转；3)随机裁剪，但裁剪尺寸大于原始图像尺寸的1/2。

3.3　评估指标

本文使用准确率(P)、召回率(R)和F-measure(F)对算法进行评估，具体定义如下：

$$P=\frac{T_{\mathrm{TP}}}{T_{\mathrm{TP}}+F_{\mathrm{FP}}} \tag{10}$$

$$R=\frac{T_{\mathrm{TP}}}{T_{\mathrm{TP}}+F_{\mathrm{FN}}} \tag{11}$$

$$F=2\times\frac{P\times R}{P+R} \tag{12}$$

其中：TP表示将正样本预测为正样本数目；FP表示将负样本预测为正样本的误报数；FN表示将正样本预测为负样本的漏报数目。准确率(P)与召回率(R)之间可能会出现矛盾的情况，其中一个测试指标较高，而另外一个测试指标较低。这时就需要综合考虑两者指标的情况，采取F-measure评估方法。

3.4　消融实验

3.4.1　多尺度特征权重融合模型的有效性

为验证多尺度特征权重融合模型对本文多尺度场景文字检测网络性能的影响，保持主干网为ResNet50不变，通过对网络添加和去除多尺度特征权重融合模型分别进行训练。本文实验在数据集IC13和IC15上分别进行了测试。实验结果如表1所示(粗体表示最优值)，在IC13数据集上添加多尺度特征权重融合模型，相较于ResNet50+FPN网络 F 值提升2%，相较于TridentNet网络 F 值提升1%。在IC15数据集上添加多尺度特征权重融合模型(MSWF Model)，相较于ResNet50+FPN网络 F 值提升2%，相较于TridentNet网络 F 值提升1%。实验结果表明多尺度特征权重融合模型有效提升了网络的检测性能。

表1　多尺度特征权重融合模型的消融实验结果

Table 1　Ablation experiment result of multi-scale feature weight fusion model

网络结构	IC13数据集			IC15数据集		
	R	P	F	R	P	F
ResNet50+FPN	0.92	0.87	0.89	0.84	0.87	0.85
TridentNet[21]	0.93	0.88	0.90	0.85	0.88	0.86
ResNet50+MFWF	0.93	0.89	**0.91**	0.86	0.87	**0.87**

3.4.2　学习主动中心轮廓模型对网络性能的影响

研究实验结果证明，更优损失函数的使用可以提高大规模图像分类和目标检测的性能。为了更好

地验证本文提出网络的检测能力，本文实验通过使用不同的损失函数，在MSRA-TD500数据集上进行测试来验证学习主动中心轮廓模型（LACC Model）对于网络检测能力的影响。保持网络结构不变，使用不同的损失函数分别进行实验，实验结果如表2所示（粗体表示最优值），使用学习主动中心轮廓模型相较于L1 Loss和SmoothL1 Loss，F值分别提升4%和2%，说明本文提出的学习主动中心轮廓模型可以更好地检测场景文本的边界，提高检测性能。

表2 学习主动中心轮廓模型的消融实验结果

Table 2 Ablation experiment result of learning active center contour model

损失函数	R	P	F
L1 Loss	0.85	0.74	0.79
SmoothL1 Loss[33]	0.87	0.75	0.81
LACC Model	**0.88**	**0.78**	**0.83**

3.5 结果对比分析

本文主要进行了以下方面的实验：

1）多尺度文本检测实验。本文在MSRA-TD500数据集上对本网络进行了测试，以验证其检测多尺度文本的能力。本实验分别在GTX1080Ti×2个GPU上进行训练和测试。将本文方法和其他方法进行比较，具体结果如表3所示（粗体表示最优值）。在MSRA-TD500数据集上本文提出网络相较于TextSnake召回率分别提升5%；相较于Lyu et al准确率提升2%；相较于最近方法MSR和TridentNet方法F值分别提升1%和1%。由此可以得出本文网络准确率和召回率相较于大部分现有方法都有所提升，本文网络综合指标F值高于最新方法，证明了本文方法对于多尺度场景文本检测的有效性。

表3 在MSRA-TD500数据集上的检测结果

Table 3 Detection results on MSRA-TD500 dataset

方法	R	P	F
SegLink(2017)[34]	0.87	0.70	0.77
East(2017)[17]	0.87	0.67	0.76
PixelLink(2018)[18]	0.84	0.73	0.78
TextSnake(2018)[35]	0.83	0.74	0.78
RRD(2018)[36]	0.87	0.73	0.79
Lyu et al(2018)[37]	0/88	0.76	0.81
MSR(2019)[15]	0.87	0.77	0.82
TridentNet(2019)[21]	0.86	0.78	0.82
本文方法	**0.88**	**0.78**	**0.83**

2）多方向英文文本检测实验。本文在IC13和IC15数据集上对本文方法进行了测试，以验证其检测多方向英文文本的能力。实验采用本文在IC17MLT上的训练模型作为预训练模型，分别在GTX1080Ti×2个GPU上进行训练和测试。将本文方法和其他方法进行比较，具体结果如表4所示（粗体表示最优值）。实验结果表明，在IC13数据集上本文提出网络相较于CTPN、TextBoxss和SSTD，召回率分别提升10%、10%和7%；相较于R2CNN，准确率提升6%；相较于最近方法PixelLink和TridentNet方法，F值分别提升2%和1%。由此可以得出本文方法召回率和准确率相较于大部分现有方法有所提升。但是本文方法准确率相较于CTPN并没有提升，原因在于：CTPN方法使用了长短期双向记忆模型处理文字建议序列，但本文方法综合指标F值高于最新方法，证明了本文方法的有效性。在IC15数据集上本文方法相较于CTPN和SSTD，召回率分别提升12%和6%；相较于EAST和R2CNN，准确率提升7%和7%；相较于最近方法TridentNet和PSENet，F值分别提升1%和1%。由此可以得出：本文方法准确率和召回率相较于大部分现有方法有所提升，而且本文方法综合指标F值高于最新方法，说明本文方法可以很好的检测多方向英文场景文本。

表4 在IC13和IC15数据集上的检测结果

Table 4 Detection results on IC13 and IC15 dataset

方法	IC13数据集			IC15数据集		
	R	P	F	R	P	F
CTPN(2016)[8]	0.83	**0.93**	0.88	0.74	0.52	0.61
TextBoxss(2016)[10]	0.83	0.88	0.85	—	—	—
SSTD(2017)[11]	0.86	0.88	0.87	0.80	0.74	0.77
EAST(2017)[17]	—	—	—	0.86	0.80	0.83
R2CNN(2017)[9]	**0.94**	0.83	0.88	0.85	0.80	0.83
PixelLink(2018)[18]	0.88	0.89	0.89	0.83	0.82	0.82
PSENet(2019)[14]	—	—	—	0.84	0.87	0.86
TridentNet(2019)[21]	0.93	0.88	0.90	0.85	**0.88**	0.86
本文方法	0.93	0.89	**0.91**	**0.86**	0.87	**0.87**

3）多方向多语言文本检测实验。为了测试本文方法对多方向多语言场景文本检测的鲁棒性，本文实验在IC17-MLT基准数据集上对其进行了评估。在IC17MLT上使用GTX1080Ti×2个GPU进行训练和测试。与最新方法的对比如表5所示（粗体表示最优值），在IC17MLT数据集上本文提出方法相较于SCUT_DLVClab1和FOTS，召回率分别提升17%和14%；相较于AF_RPN，准确率分别提升1%；相较于最近方法TridentNet和PSENet，F值分别提升1%和2%。由此可以得出：本文方法召回率相较于大部分现有方法有所提升，但是本文方法准确率相较于FOTS并没有提升，原因在于：本文方法对于场景图片中的类文字元素，如栏杆、叶子、图标等区分能力不够高，存在误检，这也是本文下一步要改进的工作。但本文方法综合指标F值高于最新方法，表明本文方法对多语言场景文本检测的有效性，可以很好地检测多方向多语言场景文本。

表 5　在 IC17MLT 数据集上的测试结果

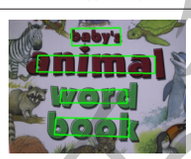
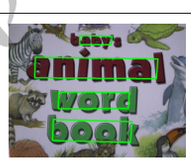
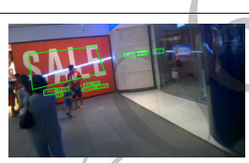
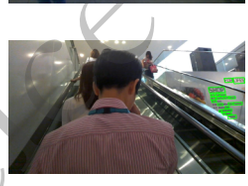
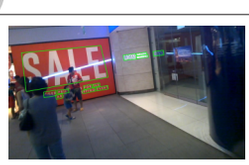
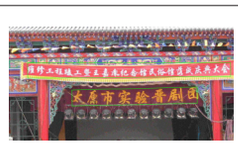
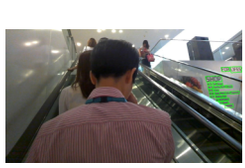
Table 5　Detection results on IC17MLT dataset

方法	*R*	*P*	*F*
SCUT_DLVClab1(2017)[30]	0.55	0.80	0.65
FOTS(2018)[38]	0.58	**0.81**	0.67
AF_RPN(2018)[39]	0.66	0.75	0.70
PSENet(2019)[14]	0.69	0.75	0.72
TridentNet(2019)[21]	0.71	0.75	0.73
本文方法	**0.72**	0.76	**0.74**

表 6 为本文方法与基于双向特征金字塔结构的 PixelLink 和三叉戟方法 TridentNet 在各数据集上代表性检测结果。从表 6 可以看出：PixelLink 和 TridentNet 方法对于某些大尺度的文本目标存在欠检测，即大尺度文本检测框不能完全包围目标，小尺度文本目标漏检的情况；而本文方法采用多尺度权重融合与学习主动中心轮廓模型相结合的方式，对于大尺度文本检测效果更好，检测框包围更准确，进一步提高小目标检测能力，有利于解决小目标漏检的问题。

表 6　不同方法在各数据集上的代表性检测结果

Table 6　Representative detect results of different methods on each dataset

方法	MSRA-TD500 数据集	IC13 数据集	IC15 数据集	IC17MLT 数据集
PixelLink(2018)[18]				
				
TridentNet(2019)[21]				
				
本文方法				
				

4　结束语

本文针对场景文本多尺度变化造成的小文本漏检、大文本欠检测以及场景文本边界检测错误问题，提出基于学习主动中心轮廓模型的场景文本检测网络。通过多尺度特征权重融合模型解决多尺度场景文本特征提取问题，基于学习主动中心轮廓模型解决场景文本边界检测错误的问题，并在 4 个公共数据集上验证了本文网络对于多尺度场景文本检测的有效性。下一步拟将本文提出网络用于弯曲场景文本检测，以提高网络的泛化性能，并研究本文网络对于类文本元素的检测能力，以增强网络的鲁棒性。

参考文献

[1] 姜维，张重生，殷绪成. 基于深度学习的场景文字检测综述[J]. 电子学报，2019，47(5)：1152-1161.
JIANG W，ZHANG C S，YIN X C. Deep learning based scene text detection：a survey[J]. Acta Electronica Sinica，2019，47(5)：1152-1161. (in Chinese)

[2] 王建新,王子亚,田萱. 基于深度学习的自然场景文本检测与识别综述[J]. 软件学报,2020,31(5):1465-1496.
WANG J X,WANG Z Y,TIAN X. Review of natural scene text detection and recognition based on deep learning[J]. Journal of Software,2020,31(5):1465-1496. (in Chinese)
[3] SINGH B,DAVIS L S. An analysis of scale invariance in object detection-SNIP[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press,2018:3578-3587.
[4] 杨玲玲,叶东毅. 一种基于图像矩和纹理特征的自然场景文本检测算法[J]. 小型微型计算机系统,2016,37(6):1313-1317.
YANG L L,YE D Y. Moment and texture based algorithm for text detection in natural scene Images[J]. Journal of Chinese Computer Systems,2016,37(6):1313-1317. (in Chinese)
[5] 尹芳,陈德运,吴锐. 改进的Gabor小波变换特征提取方法[J]. 计算机工程,2012,38(15):145-147.
YIN F,CHEN D Y,WU R. Improved Gabor wavelet transformation feature extraction method[J]. Computer Engineering,2012,38(15):145-147. (in Chinese)
[6] YAO C,BAI X,LIU W Y,et al. Detecting texts of arbitrary orientations in natural images[C]//Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press,2012:1083-1090.
[7] SHI C Z,WANG C H,XIAO B H,et al. Scene text recognition using part-based tree-structured character detection [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press,2013:2961-2968.
[8] TIAN Z,HUANG W L,HE T,et al. Detecting text in natural image with connectionist text proposal network[C]//Proceedings of European Conference on Computer Vision. Berlin,Germany:Springer,2016:56-72.
[9] JIANG Y Y,ZHU X Y,WANG X B,et al. R2CNN:rotational region CNN for orientation robust scene text detection[EB/OL]. [2021-01-05]. https://arxiv preprint arxiv:1706. 09579.
[10] LIAO M H,SHI B G,BAI X,et al. TextBoxes:a fast text detector with a single deep neural network[C]//Proceedings of AAAI Conference on Artificial Intelligence. [S. l.]:AAAI Press,2017:235-136.
[11] HE P,HUANG W L,HE T,et al. Single shot text detector with regional attention [C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C. ,USA:IEEE Press,2017:3047-3055.
[12] SZEGEDY C,LIU W,JIA Y,et al. Going deeper with convolutions [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press,2015:1-9.
[13] LIN T Y,DOLLAR P,GIRSHICK R,et al. Feature pyramid networks for object detection[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press,2017:2117-2125.
[14] WANG W H,XIE E Z,LI X,et al. Shape robust text detection with progressive scale expansion network[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press,2019:9336-9345.
[15] XUE C H,LU S J,ZHANG W. MSR:multi-scale shape regression for scene text detection[EB/OL]. [2021-01-05]. https://arxiv preprint arxiv:1901. 02596.
[16] HE T,HUANG W L,QIAO Y,et al. Accurate text localization in natural image with cascaded convolutional text network[EB/OL]. [2021-01-05]. https://arxiv preprint arxiv:1603. 09423.
[17] ZHOU X Y,YAO C,WEN H,et al. EAST:an efficient and accurate scene text detector [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press,2017:2642-2651.
[18] DENG D,LIU H F,LI X L,et al. Pixellink:detecting scene text via instance segmentation[C]//Proceedings of AAAI Conference on Artificial Intelligence. [S. 1.]:AAAI Press,2018:358-367.
[19] HE K M,ZHANG X Y,REN S Q,et al. Identity mappings in deep residual networks[C]//Proceedings of European Conference on Computer Vision. Berlin,Germany:Springer,2016:630-645.
[20] ZHU X,HU H,LIN S,et al. Deformable ConvNets V2:more deformable,better results[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press,2019:9300-9308.
[21] LI Y H,CHEN Y T,WANG N Y,et al. Scale-aware trident networks for object detection[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Washington D. C. ,USA:IEEE Press,2019:6053-6062.
[22] YU F,KOLTUN V. Multi-scale context aggregation by dilated convolutions[EB/OL]. [2021-01-05]. https://arxiv preprint arxiv:1511. 07122.
[23] WANG G R,WANG K,LIN L. Adaptively connected neural networks[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press,2019:1781-1790.
[24] KASS M,WITKIN A,TERZOPOULOS D. Snakes:active contour models[J]. International Journal of Computer Vision,1988,1(4):321-331.
[25] BRESSON X,ESEDOGLU S,VANDERGHEYNST P,et al. Fast global minimization of the active contour/snake model[J]. Journal of Mathematical Imaging and Vision,2007,28(2):151-167.
[26] LIN T Y,GOYAL P,GIRSHICK R,et al. Focal loss for dense object detection[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C. ,USA:IEEE Press,2017:2999-3007.
[27] LAW H,DENG J. CornerNet:Detecting objects as paired keypoints[C]//Proceedings of European Conference on Computer Vision. Washington D. C. ,USA:IEEE Press,2018:734-750.
[28] KARATZAS D,SHAFAIT F,UCHIDA S,et al. ICDAR 2013 robust reading competition[C]//Proceedings of the 12th International Conference on Document Analysis and Recognition. Washington D. C. ,USA:IEEE Press,2013:1484-1493.
[29] KARATZAX D,GOMEZ-BIGORDA L,NICOLAOU A,et al. ICDAR 2015 competition on robust reading[C]//Proceedings of the 13th International Conference on Document Analysis and Recognition. Washington D. C. ,USA:IEEE Press,2015:1156-1160.

(下转第262页)

（上接第252页）

[30] NAYEF N, YIN F, BIZID I, et al. ICDAR 2017 robust reading challenge on multi-lingual scene text detection and script identification-RRC-MLT[C]//Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition. Washington D. C., USA: IEEE Press, 2017: 1454-1459.

[31] DENG J, DONG W, SOCHIR R, et al. ImageNet: a large-scale hierarchical image database[C]//Proceedings of 2009 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2009: 248-255.

[32] KINGMA D P, BA J. Adam: a method for stochastic optimization[EB/OL]. [2021-01-05]. https://arxiv preprint arxiv: 1412. 6980.

[33] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks [EB/OL]. [2021-01-05]. https://arxiv preprint arxiv: 1506. 01497.

[34] SHI B, BAI X, BELONGIE S. Detecting oriented text in natural images by linking segments[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 3482-3490.

[35] LONG S B, RUAN J Q, ZHANG W J, et al. TextSnake: a flexible representation for detecting text of arbitrary shapes [C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2018: 20-36.

[36] LIAO M H, ZHU Z, SHI B G, et al. Rotation-sensitive regression for oriented scene text detection [C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 5909-5918.

[37] LYU P, YAO C, WU W, et al. Multi-oriented scene text detection via corner localization and region segmentation [C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 7553-7563.

[38] LIU X B, LIANG D, YAN S, et al. FOTS: fast oriented text spotting with a unified network[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 5676-5685.

[39] ZHONG Z Y, SUN L, HUO Q. An anchor-free region proposal network for Faster R-CNN-based text detection approaches[J]. International Journal on Document Analysis and Recognition, 2019, 22(3): 315-327.

编辑 索书志